TecnoLógicas
ISSN-p 0123-7799
ISSN-e 2256-5337
Vol. 21, No. 42, pp. 199-209
Mayo-agosto de 2018

Artículo de Investigación/Research Article

Modelo para detección automática de errores léxico-sintácticos en textos escritos en español

Model for automatic detection of lexical-syntactic errors in texts written in Spanish

María D. Bustamante-Rodríguez[1], Alberto A. Piedrahita-Ospina[2] y Iliana M. Ramírez-Velásquez[3]

Recibido: 29 de septiembre de 2017
Aceptado: 08 de marzo de 2018

Cómo citar / How to cite

M. D. Bustamante-Rodríguez, A. A. Piedrahita-Ospina, y I. M. Ramírez-Velásquez, Modelo para detección automática de errores léxico-sintácticos en textos escritos en español. *TecnoLógicas*, vol. 21, no. 42, pp. 199-209, 2018.

1 Magíster en Educación, Facultad de Ciencias Exactas y Aplicadas, Instituto Tecnológico Metropolitano, Medellín-Colombia, mariabustamante@itm.edu.co

2 Magíster en Ingeniería de Sistemas, Facultad de Ciencias Exactas y Aplicadas, Instituto Tecnológico Metropolitano, Medellín-Colombia, albertopiedrahita@itm.edu.co

3 Magíster en Automatización y Control Industrial, Facultad de Ciencias Exactas y Aplicadas, Instituto Tecnológico Metropolitano, Medellín-Colombia, ilianaramirez@itm.edu.co

Resumen

La valoración de textos escritos es una tarea que considera principalmente dos aspectos: el sintáctico y el semántico. El primero de ellos, se enfoca en la forma del texto y el segundo en el significado. La puesta en marcha de dicha tarea realizada en forma manual implica un esfuerzo en tiempo y recursos, que se puede reducir si parte del proceso se lleva a cabo de forma automática. De acuerdo con los antecedentes revisados en la corrección automática de textos, se identifican diferentes técnicas, entre ellas la lingüística, la cual se centra en los elementos sintácticos, semánticos y pragmáticos. Así, la investigación en curso se orienta a la revisión automática de textos escritos en español desde el punto de vista de la sintaxis, como punto de partida para garantizar la coherencia y la cohesión en la composición de textos, lo que puede ser de utilidad e impacto en el medio académico.

Con el propósito de llevar a cabo este estudio se recolectó y analizó un conjunto de textos de estudiantes de un programa académico, al cual se le aplicó técnicas de procesamiento de lenguaje natural y aprendizaje de máquina. Adicionalmente, se realizó una corrección manual con la finalidad de comparar ambos resultados. De esta manera, se determinó que hay correspondencia entre ellos, lo cual permitió concluir que el método automático sirve de apoyo en el proceso de corrección sintáctica de un texto escrito en español.

Palabras clave

Lingüística computacional, análisis de texto, procesamiento de lenguaje natural, inteligencia artificial, sintaxis.

Abstract

Evaluating written texts is a task that mainly considers two aspects: syntactics and semantics. The first one focuses on the form of the text, and the second one, on its meaning. Conducting this task manually implies an effort in time and resources that can be reduced if part of the process is carried out automatically. According to the reviewed literature, there are different techniques for automatically correcting texts. One of them is the linguistic approach, which focuses on syntactic, semantic, and pragmatic elements. Likewise, this ongoing research is concerned with the automatic evaluation of syntactic errors in texts written in Spanish as a starting point to ensure coherence and cohesion in text composition, which may be useful in the academic environment. In order to carry out this study, a set of texts by students enrolled in an academic program was collected and analyzed by applying natural language processing and machine learning techniques. Additionally, the content of the corpus was manually corrected to compare the results of both methods, and correspondence was established between them. For this reason, it was concluded that the automatic method supports the syntactic correction process of a text written in Spanish.

Keywords

Computational linguistics, text analysis, natural language processing, artificial intelligence, syntax.

1. INTRODUCCIÓN

La lingüística computacional también conocida como Procesamiento del Lenguaje Natural (PLN) es un campo interdisciplinario alienado con las áreas de la lingüística aplicada y la inteligencia artificial, el cual tiene como finalidad diseñar e implementar aplicaciones informáticas que emulen habilidades humanas a situaciones que involucran el lenguaje [1]. La inteligencia artificial se define como el estudio de los agentes que reciben percepciones del entorno, implementan funciones que estructuran las secuencias de las percepciones en acciones. Tales funciones se pueden representar de diferentes formas, como sistemas de producción, planificadores condicionales en tiempo real, redes neuronales, agentes reactivos y sistemas para la toma de decisiones. Por ello, la inteligencia artificial sintetiza y automatiza tareas intelectuales siendo relevante para cualquier ámbito de la actividad intelectual humana [2]. De otro lado, y de manera muy general, la lingüística moderna se ocupa del estudio científico de las lenguas naturales: su evolución histórica, su estructura interna y el conocimiento que los hablantes poseen de su propia lengua [3].

El campo híbrido que cobija la inteligencia artificial y la lingüística moderna es el ya mencionado, lingüística computacional o procesamiento del lenguaje natural, que estudia la manera de construir modelos del lenguaje con la finalidad de ser entendibles por los computadores [4]. El entendimiento del lenguaje va más allá del entendimiento de las sentencias, se fundamenta en su comprensión y en la de su contexto [2].

En esta línea, cabe acotar que se abordaron para este estudio los presupuestos de la ciencia del texto [5], dado que se analizan tanto las características y estructuras generales del uso de la lengua, como los contextos comunicativos en los que este sistema se pone en juego.

Se parte de la definición de texto como una unidad de significado constituida por una secuencia de proposiciones que cumplen con los criterios de cohesión y coherencia. La coherencia es la relación que se establece entre proposiciones enteras, postulados y conceptos que enmarcan el sentido del texto y le dan continuidad temática. Además, permite comprender la relación entre proposiciones independientes, la cual se evidencia en un nivel semántico y referencial. Aparte de resultar coherente, un texto debe tener cohesión, es decir, mantener conexión e hilaridad entre las partes de esas proposiciones, mediante conectores léxicos y gramaticales. Bajo estas condiciones, todo texto posee, a su vez, una macroestructura y una microestructura. La primera es de naturaleza semántica, puesto que define aquellas interrelaciones que se derivan del texto completo y permiten que este mantenga un sentido global. Mientras que la microestructura, hace referencia a la secuencia de oraciones que conforman el texto.

La microestructura, desde el punto de vista de la sintaxis, permite identificar, entre otras, las reglas para construir oraciones inteligibles, así como las posibilidades para combinarlas haciendo un uso adecuado de las categorías gramaticales. Es por ello que las estructuras oracionales constituyen una unidad fundamental de sentido que debe ser comprensible e interpretable. En esa línea, para garantizar las condiciones necesarias de cohesión y coherencia global en un texto, es indispensable especificar los errores sintácticos que se presenten tanto en oraciones como en secuencias de oraciones, durante el proceso de evaluación del mismo.

De otro lado, la creación de modelos computacionales que permitan escribir programas informáticos capaces de realizar tareas en donde interviene el lenguaje natural [6], ellos son de gran utilidad para la puesta en marcha de procesos automáticos de corrección sintáctica de textos escritos. A la inteligencia artificial le compete la

codificación de programas con facultades cognitivas, en esta línea, la lingüística computacional o PLN se encarga del tratamiento de la estructura lingüística, integrándose como módulo de entrada/salida dentro de un sistema compuesto [7].

La Lingüística Computacional (LC) se ocupa de investigar los mecanismos que posibilitan la comunicación entre las personas por medio del lenguaje, agregando el uso de las ciencias de la computación. Como parte de las aplicaciones en este campo, se encuentran la generación de discursos, recuperación de información, extracción de información, traducción automática, reconocimiento de voz, búsqueda de respuestas, entre otras. Así mismo, se utilizan diferentes modelos o métodos para llevar a cabo el proceso de lenguaje natural, entre ellos modelos probabilísticos del lenguaje, los cuales se definen como una distribución de probabilidad sobre un conjunto de cadenas de caracteres o de palabras contenidas en una colección base y vasta de textos escritos (corpus), o modelos basados en gramáticas [2]. En esta línea, se han reportado trabajos que incluyen el uso de n-gramas y su sintaxis para predecir los rasgos de edad, género y personalidad que tiene el autor de un determinado texto, dichos rasgos son denominados como etiquetas; el método que se describe tiene un enfoque de aprendizaje supervisado, donde un clasificador es entrenado independientemente para cada etiqueta; de esta manera, la predicción para una instancia es la unión de las salidas de cada clasificador, utilizan los n-gramas sintácticos como marcadores de personalidad junto con el uso del clasificador [8].

A partir del concepto n-gramas sintácticos (sn-gramas) y tomando como base los n-gramas de palabras tradicionales, otros investigadores llevaron el análisis sintáctico a los métodos de aprendizaje automático; ellos reportan que los sn-gramas se construyen siguiendo caminos en árboles sintácticos ya que los gramas vecinos son tomados siguiendo relaciones sintácticas en árboles sintácticos, y no tomando palabras como aparecen en un texto, de acuerdo con sus resultados, los sn-gramas se pueden aplicar en cualquier tarea de Procesamiento de Lenguaje Natural (PLN), reemplazando los n-gramas tradicionales; aplicaron tres clasificadores: Máquinas de Soporte Vectorial (SVM), Redes Bayesianas (NB) y J48; de los tres, los mejores resultados fueron presentados por el clasificador SVM [9].

La clasificación de documentos multilengua en redes sociales, se ha llevado a cabo a través de la implementación de un algoritmo que combina los n-gramas de caracteres y los n-gramas de etiquetas gramaticales. Cabe notar que la extracción de información estilística codificada en los documentos se realizó a partir de una normalización dinámica dependiente del contexto. El algoritmo se aplicó a dos corpus, primero el denominado "Comentarios de la Ciudad de México en el tiempo" y los tweets del corpus de entrenamiento de la tarea Author Profiling de PAN-CLEF 2015. Los resultados presentaron una exactitud cercana al 90 % [10].

Como se ha venido describiendo, en la lingüística computacional se desarrollan aproximaciones a las problemáticas de extracción de información, paráfrasis y minería de datos en textos; a través de técnicas, tales como redes neuronales artificiales, árboles de decisión y en general, algoritmos de aprendizaje supervisado. En esta ruta, se han reportado proyectos en los que se busca procesar textos por medio de técnicas de aprendizaje automático, en donde desarrollaron un conjunto de herramientas, con diversos fines, entre los que se encuentran construcción de material de entrenamiento, procesamiento de datos estructurados y detección de similitudes entre fragmentos de textos. En su totalidad, el sistema creado incluye una aplicación web que permite la manipulación de datos de diversos orígenes, tales como archivos con información proveniente de motores de bases de datos, para aplicar en

ellos técnicas de análisis de texto; una segunda aplicación se refiere a la lectura y edición de corpus, la realización del etiquetado sobre los corpus agregando información lingüística; y una tercera para la detección de similitudes [11].

Al considerar de manera más específica el concepto de Lingüística de Corpus, este encuadra como un enfoque metodológico para el estudio de las lenguas, además, representa oportunidades para la descripción y análisis de discursos, la construcción de gramáticas, diccionarios y otros, tanto de discursos generales como especializados, orales y escritos [12]. Sin embargo, en la literatura consultada, se han encontrado propuestas para describir textos desde el punto de vista semántico, en particular, han presentado el diseño de un constructor automático de modelos de dominio de conocimiento de forma automática sin corpus preexistente para describir semánticamente un contexto, tal constructor se basa en técnicas y métodos para la construcción de corpus a partir de fuentes digitales, mediante el desarrollo de librerías de software que automaticen las fases del sistema propuesto [13].

De otro lado, un texto corto, por ejemplo, un resumen, expone lo esencial de un tema específico, ha sido material para entrenar y probar modelos útiles para determinar la calidad lingüística. Otros investigadores proponen una evaluación sistemática de diversas clases de métricas a partir de la captura de varios aspectos de un texto, en este caso, un resumen. Los aspectos que tuvieron en cuenta para evaluar la calidad lingüística fueron: gramática, no redundancia, claridad referencial, enfoque, estructura y coherencia. Además, entre los factores que influyen en dicha calidad están la elección de palabras, la forma referencial de las entidades y la coherencia local. Utilizaron el clasificador SVM para marcar los resúmenes de las características definidas, las puntuaciones obtenidas en cada aspecto y característica, donde finalmente presentaron valores entre 78.5 y 92.9. El primer valor corresponde a la característica *nombre de la entidad* en el aspecto gramatical y el segundo, a la característica *continuidad* en el aspecto referencia [14].

De otro lado, se han señalado métodos de detección automática de unidades lingüísticas, de patrones léxicos o de palabras que expresan lo opuesto al sentido literal. En esta línea [15] se reporta un método para detectar de manera automática marcadores discursivos del español, dichos marcadores son elementos que establecen relaciones entre segmentos textuales con la finalidad de ordenar la lectura. Por medio de este método, el autor logró un resultado de 98 % de precisión y 97 % de cobertura. Adicionalmente, [16] propone un modelo de detección de ironía en textos escritos en español. Para su evaluación, se construyó un corpus compuesto de mensajes de *microblogging* (tweets) en español, los cuales fueron etiquetados como irónicos y no irónicos por evaluadores humanos, y en términos generales, el modelo detectó una ironía de aproximadamente 78 %.

Las aplicaciones generadas a partir de investigaciones en el área, han sido de gran apoyo para facilitar la evaluación de la calidad lingüística, sin embargo, en el ámbito educativo, se considera importante tener a la mano, una herramienta ágil como apoyo a la corrección automática de textos escritos desde el punto de vista de la sintaxis. De acuerdo a lo expuesto, se presenta una propuesta para llevar a cabo la tarea en mención, haciendo uso de las bondades permitidas por las técnicas del procesamiento de lenguaje natural, tales como la extracción de información y minería de textos basada en reglas y en aprendizaje de máquina supervisado.

2. METODOLOGÍA

Como paso inicial, se recolectaron los escritos de los estudiantes, previo consentimiento informado. Dichos escritos con-

formaron el conjunto de datos para ser analizados. Dada su naturaleza, fue necesario en primer lugar definir un corpus, el cual actúa como un repositorio de palabras y expresiones equívocas de la lengua española, los cuales son susceptibles de afectar la coherencia y la cohesión de textos escritos. Posteriormente, se implementó el modelo computacional, el cual clasifica, detecta y señala los errores presentes en el texto de acuerdo con el corpus definido. Por último, se genera un informe que contiene el número total de palabras y el número de errores señalados. Con estos datos, se calcula el cociente entre el número de errores señalados y el número total de palabras, este valor se denomina Índice de Densidad de Errores (IDE). Cabe aclarar, que este índice tiene una relación directa con problemas de coherencia y cohesión en los textos analizados.

2.1 Composición del corpus

El corpus está constituido por un listado de expresiones que no cumplen con las reglas sintácticas definidas conforme con los postulados de la gramática de la lengua española, centrado en el nivel de análisis sintáctico que corresponde a la manera en que se combinan y se disponen las oraciones y, de esta manera, determinar la ruta para la detección de errores de escritura en las relaciones sintácticas de concordancia, selección y posición [17].

2.2 Modelo léxico-sintáctico

El corpus está constituido por un listado de expresiones que no cumplen con las reglas sintácticas definidas de acuerdo con los postulados de la gramática de la lengua española, centrado en el nivel de análisis sintáctico que corresponde a la manera en que se combinan y se disponen las oraciones y de esta manera, determinar la ruta para la detección de errores de escritura en las relaciones sintácticas de concordancia, selección y posición [17].

El modelo está implementado en dos fases: la primera se encarga de etiquetar las palabras dentro del texto de acuerdo con su categoría gramatical o léxica y su función sintáctica en la oración. En esta etapa, se utilizó la librería NLTK, la cual es un kit de herramientas bajo un lenguaje de programación Python, que permite el procesamiento de lenguaje natural. Sobre esta plataforma, se desarrollaron POS Tagger (acrónimo en inglés de Part-Of-Speech Tagger), el cual, afín con los autores, se define como una pieza de software que procesa un texto escrito, y asigna a cada palabra una etiqueta que se refiere a la parte del discurso que le corresponde, tales como: sustantivo, verbo, adjetivo, entre otros. El POS Tagger consta de tres modos: etiquetar, entrenar y probar. El modo etiquetar usa un modelo pre-entrenado con aprendizaje de máquina supervisado para asignar etiquetas al texto, por su parte, el modo entrenar permite crear un nuevo modelo para etiquetar los datos que se proveen, finalmente, el modo probar permite observar qué tan correctas son las etiquetas asignadas [18], [19].

El grupo de etiquetas propuesto por [20], representa la información morfosintáctica de las palabras para varios idiomas, entre ellos el español e incluye las categorías: sustantivo, verbo, adjetivo, pronombre, determinante, artículo, adverbio, preposición y conjunción. Cada categoría se define mediante una nomenclatura que consta de atributos, valores y códigos, los cuales permiten distinguir la función morfosintáctica de cada palabra dentro de la oración. No todos los atributos correspondientes a cada categoría están definidos para el español, en ese caso se asigna el valor cero (0) en el respectivo atributo. Por ejemplo, a la entidad sustantivo se atribuye cuatro atributos: tipo (común y propio), género (masculino, femenino, neutro), número (singular y plural) y caso (nominativo, genitivo, dativo, acusativo y vocativo). Así, un sustantivo común, femenino, plural es etiquetado como N1220, el cero es asig-

nado ya que el atributo *caso* no está definido para el vocablo analizado.

El texto con las palabras etiquetadas, avanza hacia el proceso de detección sintáctica de errores (fase 2). Para ello, se construyó un corpus de expresiones que no cumplen con las reglas sintácticas, permitiendo así resaltar dentro del texto los errores que caben dentro de la categoría gramatical o léxica, los cuales repercuten en la coherencia, puesto que representan falta de claridad e imprecisión en las ideas expuestas en el texto. En cuanto a la cohesión, esta clase de errores demuestra falta de conexión entre ideas y párrafos. La Fig. 1 representa esquemáticamente el modelo.

El detector automático de errores que se exhibe en la Fig. 1, se detalla en el pseudocódigo presentado en la Fig. 2.

El proceso de detección de errores comienza con la lectura del resultado de la clasificación proveniente del texto a analizar, posteriormente se hace lectura de las expresiones contenidas en el corpus arriba definido. Para cada expresión en el corpus, se calcula el número de ocurrencias de dicha expresión en el texto, si esta cantidad es mayor a cero se cuenta como error, el dato se acumula y se resalta la palabra o expresión. Adicionalmente, se hace conteo del número total de palabras en el texto y calcula el IDE.

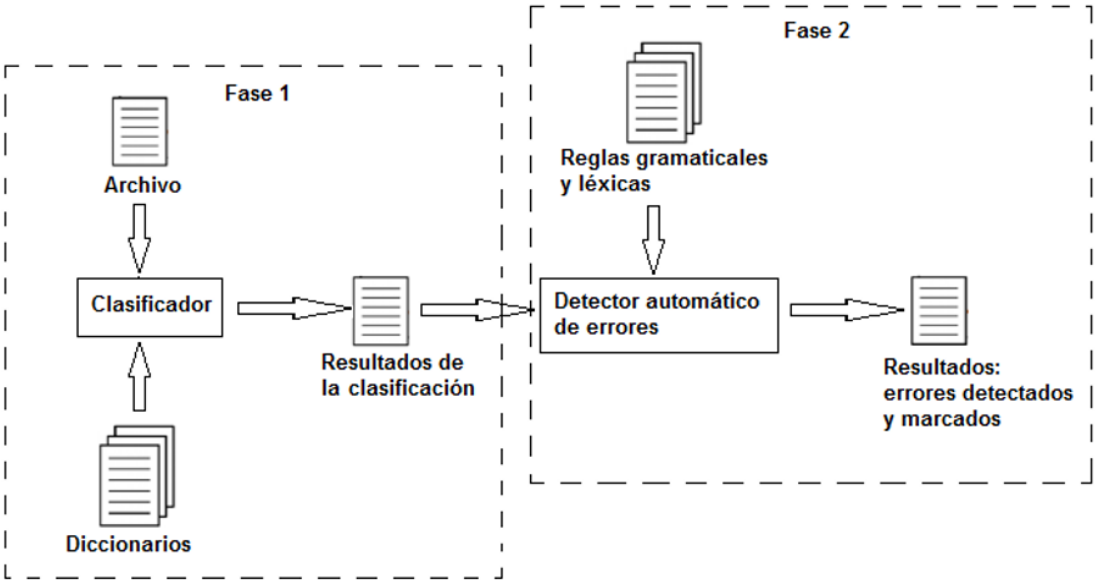


Fig. 1. Arquitectura del modelo léxico-sintáctico. Fuente: autores.

```
0   INICIO
1         Leer texto
2         Leer numero_de_palabras
3         total_expresiones=0
4         Para cada expresion_errada
5               n_ocurrencias=ocurrencias(expresion_errada, texto)
6               Si n_ocurrencias > 0 entonces
7                     resaltar_en_texto(expresion_errada, texto)
8                     total_expresiones = total_expresiones + n_ocurrencias
9               Fin si
10        Fin para
11        Mostrar total_expresiones
12        Mostrar numero_de_palabras
    FIN
```

Fig. 2. Pseudocódigo que representa el detector de errores. Fuente: autores.

2.3 Validación del modelo

Se trabajó con un conjunto de diez textos escritos, donde cada uno consta entre 52 y 183 palabras. Los escritos han sido elaborados por estudiantes de un programa académico de posgrado, cuya primera lengua es el español. El aplicativo trabaja con archivos de entradas que constituyen el corpus, el cual está compuesto palabras, oraciones y secuencias de oraciones. La herramienta señala una palabra o secuencia de palabras, de acuerdo con la existencia o no de errores sintácticos, los cuales impactan la coherencia y cohesión de un texto, puesto que al tener estos errores, la oración como unidad mínima de sentido, no es inteligible. Adicionalmente, brinda la posibilidad de declarar diversos tipos de expresiones que se ajustan a categorías gramaticales, tales como clases o subclases de palabras, artículos, pronombres, adverbios, preposiciones, conjunciones, los verbos ser y haber, así mismo, las categorías léxicas como son sustantivos, adjetivos, la mayoría de los verbos y adverbios.

Los textos recolectados fueron evaluados de manera manual para detectar errores en el campo de la sintaxis y la semántica. Posteriormente, los mismos escritos pasaron por la herramienta con la finalidad de comparar los resultados de la revisión automática con la manual.

3. RESULTADOS Y DISCUSIÓN

De acuerdo con el procedimiento descrito, los resultados obtenidos se muestran en la Tabla 1, en la cual se exponen los IDE para los diez textos analizados.

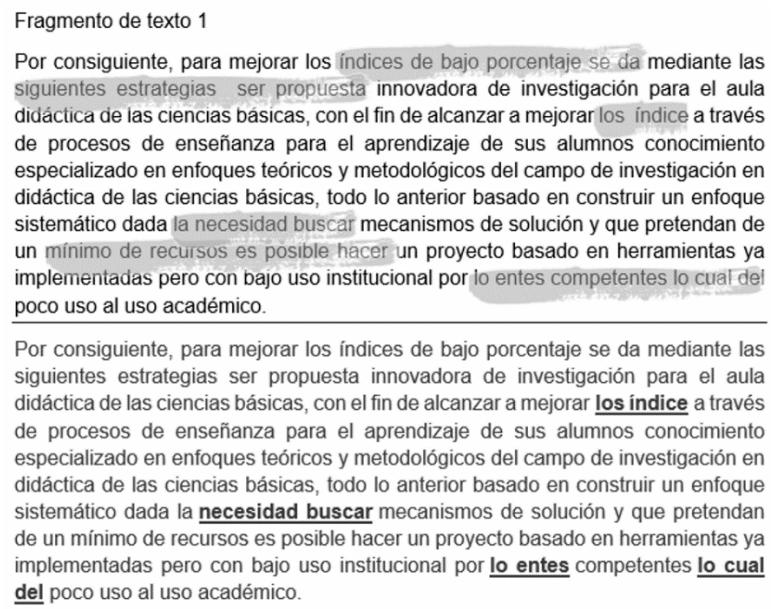
Como se puede observar en la Tabla 1, el Índice de Densidad de Errores presentado es información relevante al momento de calificar un texto desde el punto de vista de la sintaxis, por lo tanto, se verifica que el aplicativo es una herramienta de apoyo para llevar a cabo el proceso de evaluación, debido a que resalta los errores y calcula el IDE como insumo local para establecer problemas globales en la coherencia y cohesión de los textos analizados.

Para mostrar la comparación entre el método manual y el método automático de la detección de errores, se consideraron tres textos que ejemplifican la correspondencia entre el IDE y las fallas con respecto a la coherencia y la cohesión del texto. Las Fig. 3, 4 y 5 muestran los errores sintácticos marcados por el docente en la parte superior y la parte inferior presenta los marcados por el aplicativo. De acuerdo con la revisión de los resultados de la detección por ambos métodos, se puede afirmar en primera instancia, que la herramienta es consistente para la detección automática de errores sintácticos de un conjunto textos, dado que mantiene similitud con la evaluación manual de los mismos.

Tabla 1. Índice de Densidad de Errores de cada uno de los textos
Fuente: autores.

No.	Número de palabras	Índice de densidad de errores
1	119	0,034
2	157	0,038
3	84	0,071
4	183	0,032
5	84	0,036
6	154	0,019
7	94	0,010
8	85	0,047
9	78	0,025
10	52	0,038
		0,35

En los textos analizados, la herramienta identificó errores recurrentes en las relaciones sintácticas de concordancia nominal y verbal, en cuanto a número y género entre sujeto y predicado. Respecto a la relación sintáctica de selección, señaló errores como omisión de preposiciones o conjunciones y el uso innecesario de las mismas. Algunos de ellos se encuentran ejemplificados en las Fig. 3, 4 y 5.

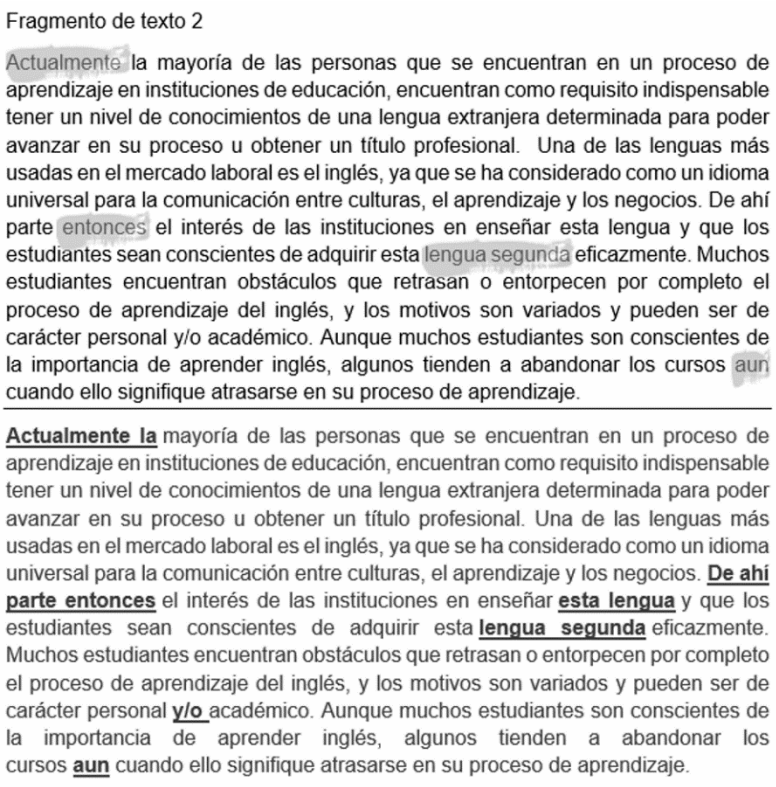


Fragmento de texto 1

Por consiguiente, para mejorar los índices de bajo porcentaje se da mediante las siguientes estrategias ser propuesta innovadora de investigación para el aula didáctica de las ciencias básicas, con el fin de alcanzar a mejorar los índice a través de procesos de enseñanza para el aprendizaje de sus alumnos conocimiento especializado en enfoques teóricos y metodológicos del campo de investigación en didáctica de las ciencias básicas, todo lo anterior basado en construir un enfoque sistemático dada la necesidad buscar mecanismos de solución y que pretendan de un mínimo de recursos es posible hacer un proyecto basado en herramientas ya implementadas pero con bajo uso institucional por lo entes competentes lo cual del poco uso al uso académico.

Por consiguiente, para mejorar los índices de bajo porcentaje se da mediante las siguientes estrategias ser propuesta innovadora de investigación para el aula didáctica de las ciencias básicas, con el fin de alcanzar a mejorar **los índice** a través de procesos de enseñanza para el aprendizaje de sus alumnos conocimiento especializado en enfoques teóricos y metodológicos del campo de investigación en didáctica de las ciencias básicas, todo lo anterior basado en construir un enfoque sistemático dada la **necesidad buscar** mecanismos de solución y que pretendan de un mínimo de recursos es posible hacer un proyecto basado en herramientas ya implementadas pero con bajo uso institucional por **lo entes** competentes **lo cual del** poco uso al uso académico.

Fig. 3. Texto muestra analizado de forma automática y manual. Fuente: autores.

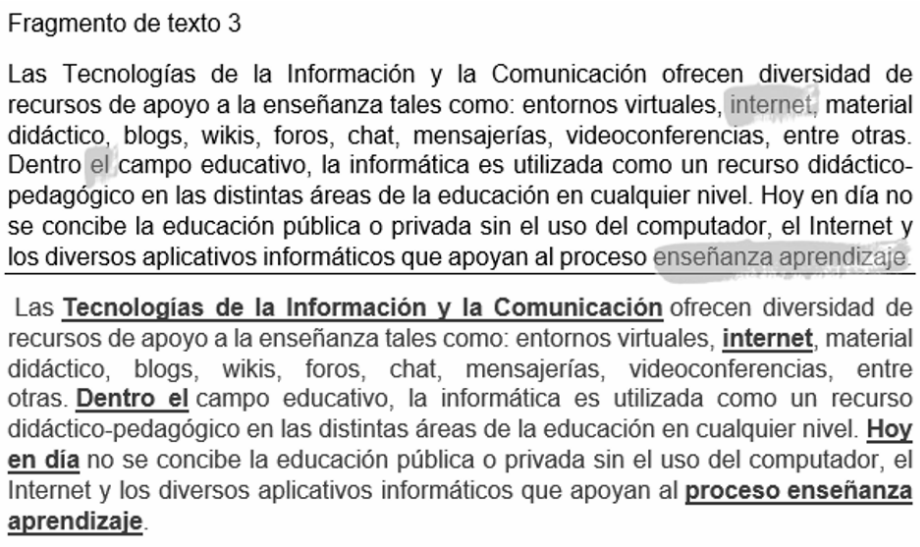


Fragmento de texto 2

Actualmente la mayoría de las personas que se encuentran en un proceso de aprendizaje en instituciones de educación, encuentran como requisito indispensable tener un nivel de conocimientos de una lengua extranjera determinada para poder avanzar en su proceso u obtener un título profesional. Una de las lenguas más usadas en el mercado laboral es el inglés, ya que se ha considerado como un idioma universal para la comunicación entre culturas, el aprendizaje y los negocios. De ahí parte entonces el interés de las instituciones en enseñar esta lengua y que los estudiantes sean conscientes de adquirir esta lengua segunda eficazmente. Muchos estudiantes encuentran obstáculos que retrasan o entorpecen por completo el proceso de aprendizaje del inglés, y los motivos son variados y pueden ser de carácter personal y/o académico. Aunque muchos estudiantes son conscientes de la importancia de aprender inglés, algunos tienden a abandonar los cursos aun cuando ello signifique atrasarse en su proceso de aprendizaje.

Actualmente la mayoría de las personas que se encuentran en un proceso de aprendizaje en instituciones de educación, encuentran como requisito indispensable tener un nivel de conocimientos de una lengua extranjera determinada para poder avanzar en su proceso u obtener un título profesional. Una de las lenguas más usadas en el mercado laboral es el inglés, ya que se ha considerado como un idioma universal para la comunicación entre culturas, el aprendizaje y los negocios. **De ahí parte entonces** el interés de las instituciones en enseñar **esta lengua** y que los estudiantes sean conscientes de adquirir esta **lengua segunda** eficazmente. Muchos estudiantes encuentran obstáculos que retrasan o entorpecen por completo el proceso de aprendizaje del inglés, y los motivos son variados y pueden ser de carácter personal **y/o** académico. Aunque muchos estudiantes son conscientes de la importancia de aprender inglés, algunos tienden a abandonar los cursos **aun** cuando ello signifique atrasarse en su proceso de aprendizaje.

Fig. 4. Texto muestra analizado de forma manual y automática. Fuente: autores.


Fragmento de texto 3

Las Tecnologías de la Información y la Comunicación ofrecen diversidad de recursos de apoyo a la enseñanza tales como: entornos virtuales, internet, material didáctico, blogs, wikis, foros, chat, mensajerías, videoconferencias, entre otras. Dentro el campo educativo, la informática es utilizada como un recurso didáctico-pedagógico en las distintas áreas de la educación en cualquier nivel. Hoy en día no se concibe la educación pública o privada sin el uso del computador, el Internet y los diversos aplicativos informáticos que apoyan al proceso enseñanza aprendizaje.

Las **Tecnologías de la Información y la Comunicación** ofrecen diversidad de recursos de apoyo a la enseñanza tales como: entornos virtuales, **internet**, material didáctico, blogs, wikis, foros, chat, mensajerías, videoconferencias, entre otras. **Dentro el** campo educativo, la informática es utilizada como un recurso didáctico-pedagógico en las distintas áreas de la educación en cualquier nivel. **Hoy en día** no se concibe la educación pública o privada sin el uso del computador, el Internet y los diversos aplicativos informáticos que apoyan al **proceso enseñanza aprendizaje**.

Fig. 5. Texto muestra analizado de forma manual y automática. Fuente: autores.

Cabe aclarar que, en algunos casos, la herramienta no señaló ningún error sintáctico, pero el evaluador humano identificó problemas de coherencia, tales como: imprecisión en las ideas expuestas, poca claridad entre las ideas principales y secundarias y falta de fluidez en la línea temática. En cuanto a problemas de cohesión, el evaluador identificó falta de conexión lógica entre ideas y párrafos.

Luego del análisis de 1090 palabras contenidas en los diez textos considerados en la prueba, el docente señaló un total de 21 errores sintácticos y el aplicativo 37. Con esta información se estimó la diferencia porcentual entre la cantidad de errores marcados entre el método automático y el manual, lo cual arrojó un valor de 76 %, es decir, el método automático supera en un 76 % al manual, en cuanto a la detección de errores en los textos mencionados evaluados por un experto en el área.

4. CONCLUSIONES

Al comparar los dos métodos, el manual y el automático, el análisis automático señaló mayor cantidad de errores sintácticos, mientras que el manual evidenció mayor identificación de errores semánticos. Esto permite inferir que la atención simultánea a elementos formales del texto, como el formato, elementos ortográficos y tipográficos, número de textos evaluados, tiempo dedicado a su lectura, entre otros, inciden en la revisión manual y pueden representar la omisión de errores sintácticos. Esto es relevante, puesto que marcar tales errores es el insumo principal para la posterior edición del estudiante, su omisión en el proceso de evaluación, puede representar dificultades para comprender falencias en criterios más amplios como la cohesión y coherencia.

En el caso de las relaciones de posición entre complemento, sujeto y posibles errores en la ubicación del verbo, la herramienta detectó menor cantidad de errores. Esto se puede explicar debido a la ambigüedad propia de los campos semánticos, así como a la presencia de otras variables en la macroestructura, como la adecuación e intención comunicativa del texto, las cuales se identifican en la evaluación manual, pues el lector conoce el contexto en el que el texto es presentado.

Dado lo anterior, se proyecta una aplicación de la herramienta que no solo detecte errores en la composición y combinación de una secuencia de oraciones, sino también, errores en expresiones de tipo discur-

sivo, errores de cohesión, tales como redundancias, repeticiones debido a la ausencia de anáforas y catáforas, así como problemas de coherencia global. Esto implicaría ampliar el corpus para el análisis, así como enriquecer el campo sintáctico y funciones de la herramienta. Adicionalmente se consideraría el análisis del caso cuando ambos sistemas no dan respuesta.

Finalmente, es posible establecer que la cantidad de tiempo requerida para la corrección sintáctica, disminuyó al utilizar la corrección automática como un apoyo en el proceso de evaluación de los textos seleccionados.

5. REFERENCIAS

[1] J. Gómez-Guinovart, "Fundamentos de lingüística computacional: bases teóricas, líneas de investigación y aplicaciones," *Bibliodoc Anu. Bibl. Doc. e Inf.*, pp. 135–146, 1998.

[2] S. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Prentice Hall, 1995.

[3] J. Corredor-Tapias and L. F. Nieto-Ruiz, "Un vistazo a los pilares de la lingüística moderna: Saussure, Chomsky y Van Dijk. Del estructuralismo a la lingüística textual," *Cuad. Lingüística Hispánica*, no. 9, pp. 83–96, 2007.

[4] G. Sidorov, *Construcción no lineal de n-gramas en la lingüística computacional*. Sociedad Mexicana de Inteligencia Artificial, 2013.

[5] T. A. Van Dijk, "Texto y Contexto. Semántica y pragmática del discurso," *Estud. Linguística Apl.*, no. 2, pp. 131–133, 1982.

[6] J. Allen, *Natural language understanding*, 2nd ed. Benjamin/Cummings Publishing Company, 1995.

[7] A. Moreno-Sandoval, *Lingüística computacional*. Madrid, España: Editorial Síntesis, 1998.

[8] J. Posadas-Durán *et al.*, "Syntactic n-grams as features for the author profiling task," *Work. Notes Pap. CLEF*, p. 5, 2015.

[9] G. Sidorov, F. Velásquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Hernández, "Syntactic N-grams as machine learning features for natural language processing," *Expert Syst. Appl.*, vol. 41, no. 3, pp. 853–860, Feb. 2014.

[10] C. González-Gallardo, J. Torres-Moreno, A. Montes-Rendón, and G. Sierra, "Perfilado de autor multilingüe en redes sociales a partir de n-gramas de caracteres y de etiquetas gramaticales," *Linguamática*, vol. 8, no. 1, pp. 21–29, 2016.

[11] J. Castillo *et al.*, "Desarrollo de sistemas de análisis de texto," in *XIX Workshop de Investigadores en Ciencias de la Computación*, 2017, pp. 58–62.

[12] G. Parodi, "Lingüística de corpus: una introducción al ámbito," *RLA. Rev. lingüística teórica y Apl.*, vol. 46, no. 1, pp. 93–119, 2008.

[13] E. A. P. Del Castillo, J. A. A. Valencia, and A. Pomares Quimbaya, "Constructor automático de modelos de dominios sin corpus preexistente," *Soc. Española para el Proces. del Leng. Nat.*, vol. 59, pp. 129–132, 2017.

[14] E. Pitler, A. Louis, and A. Nenkova, "Automatic evaluation of linguistic quality in multi-document summarization," in *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2010, pp. 544–554.

[15] W. Koza, "Marcadores discursivos del español. Descripción y propuesta de detección automática," *Rev. Epistemol. y Ciencias Humanas*, vol. 2, pp. 109–120, 2009.

[16] M. Pinto-Cruces, "Modelo de detección automática de ironía en textos en español," Universidad del Bío-Bío, 2017.

[17] Real Academia Española, *Nueva gramática de la lengua española manual*, 1st ed. Espasa, 2010.

[18] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, 2003, vol. 1, pp. 173–180.

[19] K. Toutanova and C. D. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics -*, 2000, vol. 13, pp. 63–70.

[20] G. Leech and A. Wilson, *EAGLES Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES, 1996.